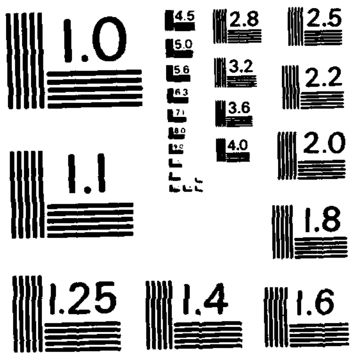
AD-A124 116 ESTIMATION IN THE PRESENCE OF NOISE OF A SIGNAL WHICH IS FLAT EXCEPT FOR..(U) MASSACHUSETTS INST OF TECH CAMBRIDGE STATISTICS CENTER Y YAO JAN 83 TR-ONR-28 1/1

UNCLASSIFIED N00014-75-C-0555 F/G 12/1 NL

END
DATE
FILMED
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS - 1963 - A

ADA 124116

STATISTICS CENTER

(12)

ESTIMATION IN THE PRESENCE OF NOISE OF A SIGNAL WHICH IS FLAT EXCEPT FOR JUMPS - PART II, THE EMPIRICAL BAYES APPROACH

BY

YI-CHING YAO

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

TECHNICAL REPORT NO. ONR 28

JANUARY 1983

PREPARED UNDER CONTRACT

N00014-75-C-0555 (NR-609-001)

FOR THE OFFICE OF NAVAL RESEARCH

Estimation in the Presence of Noise of a Signal Which is Flat Except for Jumps - Part II, The Empirical Bayes Approach

Abstract

This is the second of a two-part paper. In the first part Yao (1982), a special Bayesian Model A is studied in detail. In this part, a more general model is proposed and studied in an empirical Bayes framework. The results for Model A are applied to step-function signals using the ideas of empirical Bayes and maximum likelihood applied to the parameters of the Bayesian Model A. An efficient computational method is proposed to approximate the likelihood function under Model A. Several empirical Bayes estimators of the unknown step-function signal are compared by simulation.

Key words: Change points, nonlinear filtering, smoothing, empirical Bayes, maximum likelihood, pseudo maximum likelihood, Kullback Information.

AMS 1980 subject classification: Primary 62M20,93E14; secondary 62C12, 62G95, 93E11

1. Introduction

This is the second of a two-part paper. We consider the problem of estimating a signal which is a step function when one observes the signal plus noise. In other words, in discrete time denote the signal process by $\mu_1, \mu_2, \ldots, \mu_T$ and let $\mu_{n+1} = \mu_n$ except for occasional changes. Let the observations $X_n = \mu_n + \epsilon_n$, $1 \leq n \leq T$ where the ϵ_n are noise. We are interested in estimating μ_n based on X_i, $1 \leq i \leq T$. In the first part Yao (1982), we studied this problem in a Bayesian framework. A special Bayesian model (to be called Model A) was proposed there and the corresponding Bayes solution was derived and evaluated analytically and numerically. In the second part, we will invoke the idea of empirical Bayes to attack more general cases where not all of the assumptions of Model A are satisfied.

In the next section, a generalization of Model A is proposed. In Section 3, partial results are obtained on identifying the underlying distributions and estimating optimally the step-function signal. In Section 4, the results for Model A are applied to more general step-function signals using the ideas of empirical Bayes and maximum likelihood. In Section 5, an approximation to the likelihood is proposed and this approximation is evaluated in terms of the Kullback information. In

Section 6, several empirical Bayes estimators are studied by use of simulation.

2. A General Bayesian Model

In this section we propose the following model.

(1) The time intervals between successive changes in the signal are i.i.d.

(2) The successive heights of the signal are i.i.d.

(3) The additive noise is an i.i.d. sequence.

To be more specific,

(1') Let $\xi, \xi_1, \xi_2, \xi_3, \ldots$ be i.i.d. (F_ξ), positive integer valued with finite first moment. Let ξ' be independent of $\{\xi_n\}$ and

$$\Pr(\xi' = i) = \Pr(\xi \geq i)/E\xi \quad i = 1, 2, \ldots$$

Define the sequence of <u>change points</u> $\{n_i\}$ by

$$n_0 \equiv 0, n_1 \equiv \xi', \ n_2 \equiv \xi' + \xi_1, \ldots, n_k \equiv n' + \xi_1 + \ldots + \xi_{k-1}, \ldots$$

<u>Note:</u> The random variable ξ' is introduced in order that the 0-1 sequence generated by $\{n_1, n_2, \ldots\}$ (ones at the n_i and zeros elsewhere) be *stationary*. This is a matter of convenience and is not essential as far as asymptotic results are concerned.

(2') Let $Y, Y_0, Y_1, Y_2, \ldots$ be i.i.d. (F_Y) and represent successive heights of the signal. i.e., define the <u>signal process</u> $\{\mu_n\}$ to be

$$\mu_n \equiv Y_k \qquad \text{for} \quad n_k < n \leq n_{k+1}$$

(3') Let the <u>additive noise</u> $\varepsilon, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_T$ be i.i.d. (F_ε). Assume $E\varepsilon = 0$. Let the <u>observations</u> be

$$X_n = \mu_n + \varepsilon_n \qquad n = 1, \ldots, T$$

<u>Note:</u> Model A is a special case of this general model when F_ξ is geometrical and F_Y and F_ε are normal. Model A can be described by four parameters $p, \theta, \sigma, \sigma_\varepsilon$ where $\Pr(\xi = i) = p(1-p)^{i-1}$, $F_Y = N(\theta, \sigma^2)$ and $F_\varepsilon = N(0, \sigma_\varepsilon^{\ 2})$.

Suppose F_ξ, F_Y, and F_ε are not known. Two natural questions arise:

(Q1) Are they identifiable?

(Q2) Can μ_n be estimated "optimally"?

We designate the subsequence $\{X_n;\ i \leq n \leq j\}$ by X_i^j and we shall call estimates $\hat{\mu}_n(X_1^T, T)$ of μ_n <u>uniformly asymptotically optimal</u> relative to F_ξ, F_Y, and F_ε if

$$\lim_{T\to\infty} E_{(F_\xi,F_Y,F_\epsilon)}\{\hat{\mu}_n(X_1^T,T)-\mu_n\}^2$$

$$= \lim_{T\to\infty} E_{(F_\xi,F_Y,F_\epsilon)}\{E_{(F_\xi,F_Y,F_\epsilon)}(\mu_n|X_1^T) - \mu_n\}^2$$

uniformly for $n=1,2,\ldots,T$.

where $E_{(F_\xi,F_Y,F_\epsilon)}$ means expectation according to the probability structure determined by F_ξ, F_Y and F_ϵ. This definition is consistent in essence with that in Robbins (1964).

3. Partial Answers to (Q1) and (Q2)

Proposition 3.1 (Strong Consistency of F_ξ)

Assume that $\int x^8 dF_Y(x) < \infty$, $\int x^8 dF_\epsilon(x) < \infty$, $\mathrm{Var}(Y)>0$, and $E\xi^2 < \infty$. Then there is an estimate $\tilde{F}_\xi$ such that $\tilde{F}_\xi \to F_\xi$ w.p.1.

Proof of Proposition 3.1

Although $\{X_n\}$ is not a weakly dependent sequence, the following are still true.

$$(3.1)\quad \lim_{T\to\infty} \frac{1}{T}\sum_{n=1}^{T} X_n^\alpha = EX_1^\alpha \text{ a.s.},\quad 1 \leqq \alpha \leqq 4.$$

$$(3.2)\quad \lim_{T\to\infty} \frac{1}{T}\sum_{n=1}^{T-\ell} X_n^\alpha X_{n+\ell}^\beta = EX_1^\alpha X_{1+\ell}^\beta \text{ a.s.},\quad \ell \geqq 1,\ \alpha+\beta \leq 4.$$

The proofs of (3.1) and (3.2) appear in the Appendix.

We have

$$(3.3)\quad EX_1^\alpha X_{1+\ell}^\beta = \rho_\ell E(Y_1+\epsilon_1)^\alpha (Y_1+\epsilon_2)^\beta + (1-\rho_\ell)E(Y_1+\epsilon_1)^\alpha (Y_2+\epsilon_2)^\beta,$$

where $p_i = \Pr(\xi = i)$ and

$$(3.4)\quad \rho_\ell \equiv \Pr(\xi' \geqq 1+\ell) = \sum_{k=1+\ell}^{\infty}\ \sum_{i=k}^{\infty} p_i/E\xi$$

$$= \sum_{i=1+\ell}^{\infty} (i-\ell)p_i/E\xi$$

$$= \Big(\sum_{i=1+\ell}^{\infty} ip_i - \ell\sum_{i=1+\ell}^{\infty} p_i\Big)/E\xi$$

$$= \Big(E\xi - \sum_{i=1}^{\ell} ip_i - \ell\sum_{i=1+\ell}^{\infty} p_i\Big)/E\xi$$

$$= \Big[E\xi - \ell + \sum_{i=1}^{\ell-1} (\ell-i)p_i\Big]/E\xi\ .$$

In particular, $\rho_1 < 1$.

Substituting in (3.1) and (3.2), we have, as $T \to \infty$,

A

$$(3.5)\quad A_1 \equiv \frac{1}{T}\sum_{n=1}^{T} X_n \to EY \quad \text{a.s.}$$

$$(3.6)\quad A_2 \equiv \frac{1}{T}\sum_{n=1}^{T} X_n^2 \to EY^2 + E\varepsilon^2 \quad \text{a.s.}$$

$$(3.7)\quad A_3 \equiv \frac{1}{T}\sum_{n=1}^{T} X_n^3 \to EY^3 + 3EYE\varepsilon^2 + E\varepsilon^3 \quad \text{a.s.}$$

$$(3.8)\quad A_4 \equiv \frac{1}{T}\sum_{n=1}^{T} X_n^4 \to EY^4 + 6EY^2E\varepsilon^2 + 4EYE\varepsilon^3 + E\varepsilon^4 \quad \text{a.s.}$$

$$(3.9)\quad B_1 \equiv \frac{1}{T}\sum_{n=1}^{T-1} X_nX_{n+1} \to \rho_1 EY^2 + (1-\rho_1)(EY)^2 \quad \text{a.s.}$$

$$(3.10)\quad B_2 \equiv \frac{1}{T}\sum_{n=1}^{T-1} X_n^2X_{n+1} \to \rho_1(EY^3+E\varepsilon^2EY) + (1-\rho_1)(EY^2EY+E\varepsilon^2EY) \quad \text{a.s.}$$

$$(3.11)\quad B_3 \equiv \frac{1}{T}\sum_{n=1}^{T-1} X_n^3X_{n+1} \to \rho_1(EY^4+3E\varepsilon^2EY^2+E\varepsilon^3EY)$$

$$+ (1-\rho_1)(EY^3EY+3E\varepsilon^2(EY)^2+E\varepsilon^3EY) \quad \text{a.s.}$$

$$(3.12)\quad B_4 \equiv \frac{1}{T}\sum_{n=1}^{T-1} X_n^2X_{n+1}^2 \to \rho_1(EY^4+2E\varepsilon^2EY^2+(E\varepsilon^2)^2)$$

$$+ (1-\rho_1)((EY^2)^2+2E\varepsilon^2EY^2+(E\varepsilon^2)^2) \quad \text{a.s.}$$

Using (3.5) through (3.12) we shall show that ρ_1 can be consistently *estimated*. *First it may be seen that*

$$\Omega(EY^2) \equiv (EY^2)^2\cdot 2(A_1^2-B_1)+(EY^2)[3A_2(B_1-A_1^2)+A_1(2A_1B_1-2A_1^3+A_1A_2-B_2)$$

$$+ A_1A_3-A_2^2-B_3+B_4] + [A_1(A_3-3A_1A_2)(B_1-A_1^2)-A_1^4A_2+A_1^3B_2$$

$$- A_2^2(B_1-A_1^2)-B_1(A_1A_3-A_2^2) + (B_3-B_4)A_1^2] \to 0 \quad (T\to\infty) \quad \text{a.s.}$$

The case where $\rho_1 = 0$ is special, for then $E\zeta = 1$ and a change takes place at each time point. Then there is no way to distinguish the signal from the noise without additional information. For this reason we consider two cases:

(1) When $B_1-A_1^2 < T^{-1/3}$, estimate ρ_1 by $\hat{\rho}_1 \equiv 0$ and therefore estimate F_ζ by δ_1, the distribution with unit mass at 1.

(2) When $B_1 - A_1^2 \geq T^{-1/3}$, estimate EY^2 by the larger solution (denoted by $\hat{EY^2}$) of the quadratic equation $\Omega(EY^2) = 0$. Since $\Omega(A_1^2) = 0$, $\hat{EY}^2 \geq A_1^2$. It will be shown $\hat{EY}^2 \to EY^2$ $(T\to\infty)$ a.s. when $\rho_1 > 0$. By (3.5) and (3.9), estimate ρ_1 by $\hat{\rho}_1 \equiv \min((B_1 - A_1^2)/(\hat{EY}^2 - A_1^2), 1)$. By (3.4), estimate $E\xi$ by $\hat{E\xi} \equiv 1/(1-\hat{\rho}_1)$. By (3.2) and (3.3), estimate ρ_ℓ by $\hat{\rho}_\ell \equiv (\frac{1}{T}\sum_{n=1}^{T-\ell} X_n X_{n+\ell} - A_1^2)/(\hat{EY}^2 - A_1^2)$ for $2 \leq \ell \leq [\log T]+1$. Applying (3.4) which relates ρ_ℓ and p_ℓ, we are led to introduce $\tilde{p}_k$ recursively by

$$\tilde{p}_1 \equiv 2 - \hat{E\xi}(1-\hat{\rho}_2)$$

and

$$\tilde{p}_k \equiv \sum_{i=1}^{k-1} i\tilde{p}_i + (k+1)(1 - \sum_{i=1}^{k-1}\tilde{p}_i) - \hat{E\xi}(1-\hat{\rho}_{k+1}) \quad k = 2,3,\ldots,[\log T]$$

Let us estimate $p_1, p_2, \ldots$ by

$$\hat{p}_1 \equiv \min(\max(\tilde{p}_1, 0), 1)$$

$$\hat{p}_k \equiv \max(\min(\sum_{i=1}^{k-1}\hat{p}_i + \tilde{p}_k, 1) - \sum_{i=1}^{k-1}\hat{p}_i, 0), \quad k=2,3,\ldots,[\log T]$$

$$\hat{p}_{[\log T]+1} \equiv 1 - \sum_{i=1}^{[\log T]}\hat{p}_i$$

$$\hat{p}_k \equiv 0, \quad k > [\log T] + 1.$$

Now, we show that the estimate of F_ξ is consistent. There are two different cases:

(i) If $\Pr(\xi_1 = 1) = 1$, i.e. $\rho_1 = 0$, then

$$A_1^2 \to (EY)^2 \quad (T\to\infty) \quad \text{a.s.}$$

$$B_1 \to \rho_1 EY^2 + (1-\rho_1)(EY)^2 = (EY)^2 \quad (T\to\infty) \quad \text{a.s.}$$

Since $\{X_n\}$ is i.i.d., we can apply the law of the iterated logarithm (Philipp and Stout (1975), p. 26),

$$A_1^2 - B_1 = O(\sqrt{\log\log T/T}) \quad \text{a.s.}$$

Thus, $B_1 - A_1^2 < T^{-1/3}$ for T large enough. So $\hat{F}_\xi = \delta_1$ for T large enough. This proves the consistency when $F_\xi = \delta_1$.

(ii) If $\Pr(\xi_1 = 1) < 1$, i.e. $\rho_1 > 0$, then

$$\lim_{T\to\infty} A_1^2 = (EY)^2 < \rho_1 EY^2 + (1-\rho_1)(EY)^2 = \lim_{T\to\infty} B_1 \quad \text{a.s.}$$

using the assumption $\mathrm{Var}(Y) > 0$. So $B_1 - A_1^2 > T^{-1/3}$ for

T large enough. Since $\Omega(EY^2) \to 0$ $(T\to\infty)$ and $\Omega(A_1^{\ 2}) = 0$ and A_1^2 is bounded away from (and below) EY^2 for T large enough, $\hat{EY}^2 \to EY^2$ $(T\to\infty)$ a.s. So, from (3.9), $\hat{\rho}_1 \equiv \min((B_1 - A_1^{\ 2})/(\hat{EY}^2 - A_1^{\ 2}), 1) \to \rho_1 (T\to\infty)$ a.s. Also, from (3.4), $\hat{E\xi} \equiv 1/(1-\hat{\rho}_1) \to E\xi$ $(T\to\infty)$ a.s.

Now, we will show by induction that for all $k \geq 1$, $\tilde{p}_k \to p_k$ $(T\to\infty)$ a.s. From (3.2) with $(\alpha,\beta) = (1,1)$

$$\lim_{T\to\infty} \frac{1}{T} \sum_{n=1}^{T-\ell} X_n X_{n+\ell} = \rho_\ell E(Y_1+\varepsilon_1)(Y_1+\varepsilon_2) + (1-\rho_\ell)E(Y_1+\varepsilon_1)(Y_2+\varepsilon_2)$$

$$= \rho_\ell (EY^2 - (EY)^2) + (EY)^2 \quad \text{a.s.}$$

So, $\hat{\rho}_\ell \equiv [\frac{1}{T} \sum_{n=1}^{T-\ell} X_n X_{n+\ell} - A_1^{\ 2}]/(\hat{EY}^2 - A_1^{\ 2}) \to \rho_\ell$ $(T\to\infty)$ a.s.

Now, suppose $\tilde{p}_i \to p_i$ $(T\to\infty)$ a.s. for $i=1,2,\ldots,k-1$. From (3.4),

$$\rho_{k+1} = 1 - [k+1 - \sum_{i=1}^{k-1} (k+1-i)p_i - p_k]/E\xi$$

So,

$$\tilde{p}_k \equiv \sum_{i=1}^{k-1} i\tilde{p}_i + (k+1)(1 - \sum_{i=1}^{k-1} \tilde{p}_i) - \hat{E\xi}(1 - \hat{\rho}_{k+1}) \to p_k \ (T\to\infty) \text{ a.s.}$$

Finally, since $\lim_{T\to\infty} \tilde{p}_k = p_k$ a.s. $k=1,2,\ldots,(\hat{p}_1,\hat{p}_2,\ldots)$ is a consistent estimate of F_ξ. □

Having established Proposition 3.1, it is important to note that with the use of Equations (3.5) - (3.12) we can show that EY^2, EY^3, EY^4, $E\varepsilon^2$, $E\varepsilon^3$, and $E\varepsilon^4$ can be consistently estimated when $\rho_1 > 0$.

To approach question Q2, we derive

<u>Proposition 3.2</u>

<u>Assume</u> <u>that</u> $E\xi^2 < \infty$, $Pr(\xi=1) < 1$, $F_Y = N(\theta,\sigma^2)$ <u>and</u> $F_\varepsilon = N(0,\sigma_\varepsilon^{\ 2})$ <u>where</u> F_ξ, θ, σ^2 <u>and</u> σ_ε^2 <u>are</u> <u>unknown</u>. <u>Then</u> <u>for</u> <u>all</u> $k \geq 1$ <u>there</u> <u>exist</u> $g_n^{(k)}(X_1^T, T)$ <u>such</u> <u>that</u>

$$\lim_{T\to\infty} E(g_n^{(k)}(X_1^T,T) - \mu_n)^2 = E(E(\mu_n | X_{n-k}^{n+k}) - \mu_n)^2$$

uniformly for $n=k+1,\ldots,T-c(T)$

<u>where</u> $c(T)$ <u>is</u> <u>an</u> <u>arbitrary</u> <u>positive</u> <u>integer</u> <u>valued</u>, <u>increasing</u> <u>unbounded</u> <u>function</u>.

<u>Proof of Proposition 3.2</u>

From Proposition 3.1 there exist consistent estimates $\hat{F}_\xi(X_1^{c(T)}, c(T))$, $\hat{\theta}(X_1^{c(T)}, c(T))$, $\hat{\sigma}(X_1^{c(T)}, c(T))$ and $\hat{\sigma}_\varepsilon(X_1^{c(T)}, c(T))$ of F_ξ, θ, σ and σ_ε, respectively. In particular, $\hat{\theta}$ can be chosen to be $\sum_{i=1}^{c(T)} X_i/c(T)$. By stationarity,

$$E(\hat{E}_n(\mu_n \mid X_{n-k}^{n+k}) - \mu_n)^2 = E(\hat{E}_{k+1}(\mu_{k+1} \mid X_1^{2k+1}) - \mu_{k+1})^2$$

for $n=k+1,\ldots,T-c(T)$

where $\hat{E}_n$ = expectation under $\hat{F}_\xi(X_{n+1}^{n+c(T)}, c(T))$, $\hat{\theta}(X_{n+1}^{n+c(T)}, c(T))$, $\hat{\sigma}(X_{n+1}^{n+c(T)}, c(T))$ and $\hat{\sigma}_\epsilon(X_{n+1}^{n+c(T)}, c(T))$.

Therefore, we need only show

$$\lim_{T\to\infty} E(\hat{E}_{k+1}(\mu_{k+1} \mid X_1^{2k+1}) - \mu_{k+1})^2 = E(E(\mu_{k+1} \mid X_1^{2k+1}) - \mu_{k+1})^2.$$

Obviously,

$$\hat{E}_{k+1}(\mu_{k+1} \mid X_1^{2k+1}) \to E(\mu_{k+1} \mid X_1^{2k+1}) \quad (T\to\infty) \quad \text{a.s.}$$

Since $\hat{E}_{k+1}(\mu_{k+1} \mid X_1^{2k+1})$ is bounded by $\max(|\hat{\theta}|, |X_1|, \ldots, |X_{2k+1}|)$,

$$\lim_{T\to\infty} E(\hat{E}_{k+1}(\mu_{k+1} \mid X_1^{2k+1}) - \mu_{k+1})^2 = E(E(\mu_{k+1} \mid X_1^{2k+1}) - \mu_{k+1})^2$$

by the dominated convergence theorem. □

Remark 1:

This proof of Proposition 3.2 almost establishes the uniform asymptotic optimality of $\hat{E}_n(\mu_n \mid X_{n-k}^{n+k})$ as an estimate of μ_n for $k+1 \le n \le T-c(T)$ and k large.

Remark 2:

Although in Proposition 3.2 F_Y and F_ξ are required to be Gaussian, the same result can be proved if they belong to (regular) parametric families of distributions whose parameters can be estimated consistently.

4. An Empirical Bayes Estimate Using Model A with Unknown Parameters

In general, a step-function signal can be either deterministic or stochastic and therefore Model A, or even the general model, can fail to be satisfied. Why then should we consider these models? The basic idea is that it is hoped the unknown signal would resemble a "typical" realization of these models with properly assigned parameters or distributions. Indeed, this is a possible interpretation of the empirical Bayes idea. The most famous example is the James-Stein estimate which shows some superiority to the classical estimate of the mean of a multivariate normal distribution.

It is almost impossible to produce a sensible estimate of the signal without any information about the structure of the signal and/or the noise. Hence, our first assumption is that the noise is Gaussian white noise. One main reason to have the Gaussian assumption is that it is hard to distinguish outliers from jumps if the noise has a heavy tailed distribution. Futhermore, if the step-function

signal has many jumps, the noise variance cannot be well estimated. Indeed, the noise variance in Model A is not identifiable without further information. For instance, the observation process $\{X_n\}$ is i.i.d. $N(0,1)$ when $(p,\theta,\sigma,\sigma_\varepsilon) = (1,0,1,0)$ or $(p,0,0,1)$. So, we make the second assumption that the rate of jump in the signal is at most p_0 where p_0 is a specified number between 0 and 1.

As the next step in generalizing our estimation procedure, let us assume that Model A applies with unknown parameters $p,\theta,\sigma,\sigma_\varepsilon$ and apply maximum likelihood to estimate these parameters. To be more precise, we estimate the signal μ_n as follows. First, fit Model A to the observations $X_i (1 \le i \le T)$ by finding the maximum likelihood estimates (MLE) $\hat{p},\hat{\theta},\hat{\sigma}$ and $\hat{\sigma}_\varepsilon$ with the constraint that $p \le p_0$. Next, estimate μ_n by

$$(4.1)\quad \hat{\mu}_n^{EB} \equiv E_{(\hat{p},\hat{\theta},\hat{\sigma},\hat{\sigma}_\varepsilon)}(\mu_n \mid X_1^T)$$

where EB stands for empirical Bayes.

Since the MLE satisfy, (for constants $a \neq 0, c$),

$$\hat{p}(aX_1+c,\ldots,aX_T+c) = \hat{p}(X_1,\ldots,X_T)$$

$$(4.2)\quad \hat{\theta}(aX_1+c,\ldots,aX_T+c) = a\hat{\theta}(X_1,\ldots,X_T) + c$$

$$\hat{\sigma}(aX_1+c,\ldots,aX_T+c) = |a|\hat{\sigma}(X_1,\ldots,X_T)$$

$$\hat{\sigma}_\varepsilon(aX_1+c,\ldots,aX_T+c) = |a|\hat{\sigma}_\varepsilon(X_1,\ldots,X_T)$$

and since Model A is time reversible, we have

Proposition 4.1

The empirical Bayes estimator of μ_n, $\hat{\mu}_n^{EB}$, is translation invariant, scale invariant and time reversible. That is,

$$\hat{\mu}_n^{EB}(aX_1+c,\ldots,aX_T+c) = a\,\hat{\mu}_n^{EB}(X_1,\ldots,X_T) + c$$

$$\hat{\mu}_n^{EB}(X_1,\ldots,X_T) = \hat{\mu}_{T-n+1}^{EB}(X_T,X_{T-1},\ldots,X_1).$$

The computation of the MLE can be very time-consuming. A naive method may require $O(2^T)$ operations to compute the likelihood for each quadruple $(p,\theta,\sigma,\sigma_\varepsilon)$. We present in Proposition 4.2 a representation of the likelihood function which reduces the number of operations to the order of T^2. Since the log likelihood $L(p,\theta,\sigma,\sigma_\varepsilon;X_1^T)$ satisfies

(4.3) $L(p,\theta,\sigma,\sigma_\varepsilon;X_1^T) = L(p,0,\sigma/\sigma_\varepsilon,1;(X')_1^T) - T\log\sigma_\varepsilon$

where $X'_n = (X_n-\theta)/\sigma_\varepsilon$, we need only consider $L(p,0,\sigma,1;X_1^T)$.
Let $S_0 \equiv 0$ and $S_n \equiv \sum_{k=1}^{n} X_k$ for $1 \le n \le T$.

Proposition 4.2

$$L(p,0,\sigma,1;X_1^T = x_1^T) = \log f_{X_1}(x_1) + \sum_{n=1}^{T-1} \log f_{X_{n+1}}(x_{n+1}|X_1^n=x_1^n)$$

where

$$\mathcal{L}(X_1) = N(0,\sigma^2+1),$$

(4.4) $$\mathcal{L}(X_{n+1}|X_1^n) = (1-p)\sum_{k=1}^{n} A_k^{(n)}\cdot N\left(\frac{S_n-S_{n-k}}{k+\sigma^{-2}}, \frac{1}{k+\sigma^{-2}}+1\right) + p\,N(0,\sigma^2+1)$$

and $A_k^{(n)}$ are defined in Proposition 4.2 of Yao (1982).

Proof of Proposition 4.2

We need only derive (4.4). However, this is a simple consequence of Proposition 4.2 of Yao (1982) and the following identity.

$$\mathcal{L}(X_{n+1}|X_1^n) = \mathcal{L}(\mu_{n+1}+\varepsilon_{n+1}|X_1^n)$$

$$= \mathcal{L}(\mu_{n+1}|X_1^n)\circledast N(0,1)$$

$$= [(1-p)\mathcal{L}(\mu_n|X_1^n) + p\,N(0,\sigma^2)]\circledast N(0,1)$$

$$= (1-p)\mathcal{L}(\mu_n|X_1^n)\circledast N(0,1) + p\,N(0,\sigma^2+1)$$

where $\mathcal{L}_1 \circledast \mathcal{L}_2 \equiv$ the convolution of law $\mathcal{L}_1$ with $\mathcal{L}_2$. □

5. An Approximation to the Likelihood and the Pseudo MLE

Even though Proposition 4.2 suggests a way to compute the likelihood with $O(T^2)$ operations, it is still time-consuming to compute the MLE without further reduction in computation. Therefore it is desired to find a more efficient way to approximate the likelihood. We will make use of an idea of Harrison and Stevens (1976) to develop an approximation procedure which reduces the number of operations to the order of T. This idea has been used and justified in Yao (1982).

We approximate $\mathcal{L}(X_{n+1}|X_1^n)$ as follows. Again, assume $\theta = 0$ and $\sigma_\varepsilon = 1$ for simplicity. In Section 5 of Yao (1982), $N(\theta_n,\tau_n^2)$ is introduced to approximate $\mathcal{L}(\mu_n|X_1^n)$ where θ_n and τ_n^2 are defined recursively. Since

$$\mathcal{L}(X_{n+1}|X_1^n) = (1-p)N(\mu_n|X_1^n)\circledast N(0,1) + p\,N(0,\sigma^2+1)$$

we are naturally led to approximate $\mathcal{L}(X_{n+1}|X_1^n)$ by $(1-p)\,N(\theta_n,\tau_n^2+1) + p\,N(0,\sigma^2+1)$.

Now we can approximate the log likelihood $L(p,\theta,\sigma,\sigma_\varepsilon;X_1^T)$ by use of Proposition 4.2 and the above approximation and denote this approximate log likelihood by $\tilde{L}(p,\theta,\sigma,\sigma_\varepsilon;X_1^T)$.

It should be noted that this approximation is exact

when p is 0 or 1, for Model A is a Gaussian system when p is 0 or 1. Now we propose to measure this approximation in terms of the Kullback information between $\exp(L)$ and $\exp(\tilde{L})$ under Model A. More precisely, we will treat

$$(5.1)\quad I_K(p,\theta,\sigma,\sigma_\varepsilon) \equiv E_{(p,\theta,\sigma,\sigma_\varepsilon)}[L(p,\theta,\sigma,\sigma_\varepsilon;X_1^T)-\tilde{L}(p,\theta,\sigma,\sigma_\varepsilon;X_1^T)]$$

as a measure of how well L is approximated by $\tilde{L}$. Note that

$$(5.2)\quad I_K(p,\theta,\sigma,\sigma_\varepsilon) = I_K(p,0,\sigma/\sigma_\varepsilon,1)$$

We considered 63 cases where

$p\varepsilon\{0.02, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8\}$, $\sigma\varepsilon\{0.5, 1,2,3,4,5,7,10,15\}$, $\theta = 0$ and $\sigma_\varepsilon = 1$. The I_K were estimated by use of simulation with a computer (HP 3000) where 400 samples of size $T = 20$ were generated for each case. The results are presented in Table 5.1.

According to Table 5.1, $E(L-\tilde{L}) \leq 0.14$, and $SD(L-\tilde{L}) \leq 0.48$. Here $SD(Y)$ is the standard deviation of random variable Y. So,

$$-1.44 \leq E(L-\tilde{L}) - 3\,SD(L-\tilde{L}) < E(L-\tilde{L}) + 3\,SD(L-\tilde{L}) \leq 1.58$$

The probability that the likelihood ratio $\exp(L-\tilde{L})$ satisfies

$$0.24 \doteq \exp(-1.44) < \exp(L-\tilde{L}) < \exp(1.58) \doteq 4.85$$

is very high in the worst case under Model A. This suggests that the approximation will yield reasonably good results.

We shall define the pseudo MLE $\hat{p}',\hat{\theta}',\hat{\sigma}',\hat{\sigma}_\varepsilon'$ as the values of the parameters which maximize $\tilde{L}$ subject to $p \leq p_0$. Then we estimate μ_n by

$$(5.3)\quad \hat{\mu}_n' \equiv E_{(\hat{p}',\hat{\theta}',\hat{\sigma}',\hat{\sigma}_\varepsilon')}(\mu_n|X_1^T).$$

6. Simulation on Empirical Bayes Estimators

In the last section, we have introduced, for the sake of computation, $\hat{\mu}_n'$ which is an approximation to $\hat{\mu}_n^{EB}$. In order to evaluate the performance of $\hat{\mu}_n'$, we carried out the following computer simulations on an HP 3000.

We considered 21 deterministic signal sequences $\{\mu_n^{(i)}\}$ of length $T = 20$ ($1\leq n\leq 20$, $1\leq i\leq 21$). For each signal sequence, we generated 100 samples of Gaussian white noise of variance 1.

In defining $\hat{\mu}_n'$, we estimated the parameters of Model A by use of pseudo maximum likelihood. It is interesting to see how well the method of moments can do compared to the pseudo maximum likelihood method. It is also interesting to see how much the additional information $\sigma_\varepsilon = 1$ can contribute to estimating μ_n.

Hence, we considered the following four estimators of μ_n.

(i) Estimator 1 - $\hat{\mu}_n'$, $p_0 = 0.2$

(ii) Estimator 2 - This is defined in the same way as Estimator 1 except with one more constraint $\sigma_\epsilon = 1$ in the pseudo maximum likelihood estimation of the parameters.

(iii) Estimator 3 - $E_{(p_1,\theta_1,\sigma_1,\sigma_{\epsilon 1})}(\mu_n|X_1^T)$ where $p_1 = \max(\min(p_2, 0.2), 0)$, $\theta_1 = \bar{X}$ (the *sample* mean), $\sigma_1 = \max(\sigma_2, 0)$, $\sigma_{\epsilon 1} = \max(\sigma_{\epsilon 2}, 0)$ and $p_2, \sigma_2, \sigma_{\epsilon 2}$ satisfy

$$\sum_{n=1}^{T} X_n^2/T = \bar{X}^2 + \sigma_2^2 + \sigma_{\epsilon 2}^2$$

$$\sum_{n=1}^{T-1} X_n X_{n+1}/(T-1) = \bar{X}^2 + (1-p_2)\sigma_2^2.$$

$$\sum_{n=1}^{T-2} X_n X_{n+2}/(T-2) = \bar{X}^2 + (1-p_2)^2 \sigma_2{}^2.$$

(iv) Estimator 4 - $E_{(p_3,\theta_3,\sigma_3,\sigma_{\epsilon 3})}(\mu_n|X_1^T)$ where

$p_3 = \max(\min(p_4, 0.2), 0)$, $\theta_3 = \bar{X}$, $\sigma_3 = \max(\sigma_4, 0)$, $\sigma_{\epsilon 3} = 1$ and p_4, σ_4 satisfy

$$\sum_{n=1}^{T} X_n^2/T = \bar{X}^2 + \sigma_4^2 + 1$$

$$\sum_{n=1}^{T-1} X_n X_{n+1}/(T-1) = \bar{X}^2 + (1-p_4)\sigma_4^2$$

We use the average of mean squared errors (AMSE) as the criterion. The simulation results are presented in Table 6.1 where we also present the mean and standard deviation of σ_ϵ', the pseudo MLE of σ_ϵ.

<u>Note:</u>

All the four estimators have one common property. That is, they first estimate $p, \theta, \sigma, \sigma_\epsilon$ and then estimate μ_n by the corresponding Bayes estimate $E_{(p,\theta,\sigma,\sigma_\epsilon)}(\mu_n|X_1^T)$. In the simulation above, we actually computed the approximate Bayes estimate (see Yao (1982), Section 5) instead of the exact one.

<u>Remarks:</u> (Based on Table 6.1)

(1) Roughly speaking, when the number of jumps increases, the AMSE of $\hat{\mu}_n'$ increases. When the size of jumps increases, the AMSE of $\hat{\mu}_n'$ first increases and then decreases. For when the size of jumps is moderate (i.e.

compatiable with the noise) it is hard to tell where jumps take place and to take appropriate action. This property is similar to that of the Bayes estimator. (See Remark 1 of Section 7 in Yao (1982)).

(2) Estimator 1 $(\hat{\mu}_n{}')$ is better than Estimator 3. This implies that the method of pseudo maximum likelihood is significantly better than the method of moments in finding suitable parameter values.

(3) Estimator 1 is just slightly worse than Estimator 2. So the information about the noise variance is not very important for estimating the signal unless the rate of change in the signal is high. In that case, it is hard to estimate σ_ε well.

(4) The empirical Bayes estimator, $\hat{\mu}_n{}'$, is robust against the signals' behavior. However, it is not known how to deal with cases involving non-Gaussian noise which may introduce outliers under the veil of jumps.

(5) When the prior information, the rate of change $\leq p_0$, is not correct, $\hat{\mu}_n{}'$ may be misleading, although our limited simulations do not indicate so.

(6) It is interesting that $\hat{\sigma}_\varepsilon{}'$, the pseudo MLE of σ_ε, estimates σ_ε well with small bias. This is essentially due to the information $p \leq p_0$.

References

[1] Harrison, P. J. and Stevens, C.F. (1976). Bayesian forecasting. *J. Roy. Stat. Soc.* B 38, 205-247 (with discussion).

[2] Philipp, W. and Stout, W. (1975). *Almost Sure Invariance Principles for Partial Sums of Weakly Dependent Random Variables.* Amer. Math. Soc. Mem. No. 161.

[3] Robbins, H. (1964). The Empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.* 35, 1-20.

[4] Yao, Y.-C. (1982). *Estimation in the Presence of Noise of a Signal Which is Flat Except for Jumps - Part I, A Bayesian Study.* Tech. Rept. ONR 25, Statist. Center, MIT, Cambridge, MA.

Appendix

Proof of (3.1) and (3.2) in Proposition 3.1

It is not difficult to see by applying the Gaal-Koksma strong law of large numbers (Philipp and Stout (1975), Appendix 1) that we need only show

$$\lim_{T\to\infty} \frac{1}{T} \sum_{n=1}^{T} \mu_n^{\alpha} = E\,\mu_1^{\alpha} \text{ a.s.} \quad 1 \le \alpha \le 4$$

and

$$\lim_{T\to\infty} \frac{1}{T} \sum_{n=1}^{T-\ell} \mu_n^{\alpha}\mu_{n+\ell}^{\beta} = E\mu_1^{\alpha}\mu_{1+\ell}^{\beta} \quad \text{a.s. } \ell \ge 1,\ \alpha+\beta \le 4$$

For example, since

$$\frac{1}{T} \sum_{n=1}^{T} \mu_n^2 \varepsilon_n{}^2 = \frac{1}{T} \sum_{n=1}^{T} \mu_n{}^2 E\varepsilon^2 + \frac{1}{T} \sum_{n=1}^{T} \mu_n^2(\varepsilon_n^2 - E\varepsilon^2)$$

and $\lim_{T\to\infty} T^{-1} \sum_{n=1}^{T} \mu_n^2(\varepsilon_n^2 - E\varepsilon^2) = 0$ a.s. by the Gaal-Koksma strong law of large numbers, $\lim_{T\to\infty} T^{-1} \sum_{n=1}^{T} \mu_n^2\varepsilon_n^2 = E\mu_1^2 E\varepsilon^2$ a.s. if $\lim_{T\to\infty} T^{-1} \sum_{n=1}^{T} \mu_n^2 = E\mu_1^2$ a.s.

In the following we only treat one case, i.e.

$$\lim_{T\to\infty} \frac{1}{T} \sum_{n=1}^{T-\ell} \mu_n\mu_{n+\ell} = E\mu_1\mu_{1+\ell} \quad \text{a.s.} \quad \text{for } \ell \ge 1$$

and the rest can be established similarly.

$$U_n \equiv 0 \qquad , \text{ if there exists } k \text{ such that } n \le n_k < n+\ell$$

$$\equiv \mu_n\mu_{n+\ell} - EY^2 \quad , \text{ otherwise}$$

$$V_n \equiv \mu_n\mu_{n+\ell} - (EY)^2, \text{ if there exists } k \text{ such that } n \le n_k < n+\ell$$

$$\equiv 0 \qquad , \text{ otherwise}$$

$\{U_n\}$ and $\{V_n\}$ are stationary.

$$E\Big(\sum_{n=1}^{N} U_n\Big)^2 = N\,EU_1{}^2 + 2 \sum_{k=1}^{N-1} (N-k)\ EU_1U_{k+1}$$

$$= N\,EU_1{}^2 + 2 \sum_{k=1}^{N-1} (N-k)\ Pr(\xi' > k+\ell)E(Y^2-EY^2)^2$$

$$\le N\,(EU_1{}^2 + 2\,E(Y^2-EY^2)^2\,E\xi')$$

$$E\left(\sum_{n=1}^{N} V_n\right)^2 = N\,EV_1^{\,2} + 2\sum_{k=1}^{N-1} (N-k)\;EV_1V_{k+1}$$

$$= N\,EV_1^{\,2} + 2\sum_{k=1}^{\ell} (N-k)\;EV_1V_{k+1} + 2\sum_{k=\ell+1}^{N-1} (N-k)\;EV_1V_{k+1}$$

$$\leq N[EV_1^{\,2}+2\sum_{k=1}^{\ell}|EV_1V_{k+1}|+2\sum_{k=\ell+1}^{N-1}\Pr(\text{no } k_1 \text{ such that } 1+\ell\leq\eta_{k_1}<1+k)E(Y_1Y_2-(EY)^2)(Y_2Y_3-(EY)^2)]$$

$$= N[EV_1^{\,2} + 2\sum_{k=1}^{\ell}|EV_1V_{k+1}|+2\sum_{k=\ell+1}^{N-1}\Pr(\xi'>k-\ell)(EY)^2\mathrm{Var}(Y)]$$

$$\leq N[EV_1^{\,2} + 2\sum_{k=1}^{\ell}|EV_1V_{k+1}|+2\,E\xi'(EY)^2\,\mathrm{Var}(Y)]$$

Therefore, by the Gaal-Koksma strong law of large numbers,

$$\lim_{T\to\infty}\frac{1}{T}\left(\sum_{n=1}^{T-\ell} U_n + \sum_{n=1}^{T-\ell} V_n\right) = 0 \quad \text{a.s.}$$

So,

$$\lim_{T\to\infty}\frac{1}{T}\sum_{n=1}^{T-\ell} U_nU_{n+\ell} = \lim_{T\to\infty}\frac{1}{T}\,[EY^2\,c + (EY)^2(T-\ell-c)] \quad \text{a.s.}$$

where

$$c \equiv \sum_{k=1}^{d-1} \max(\eta_k-\eta_{k-1}-\ell,0) + \max(T-\ell-\eta_{d-1},0)$$

$$d \equiv \inf\{k : \eta_k > T\}$$

Now, by the strong law of large numbers,

$$\lim_{T\to\infty}\frac{T}{d} = E\xi \text{ a.s.}$$

$$\lim_{T\to\infty}\frac{c}{d} = E\max(\xi-\ell,0) \quad \text{a.s.}$$

$$= \sum_{k=1}^{\infty} k\,\Pr(\xi = k+\ell)$$

Therefore,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{n=1}^{T-\ell} \mu_n \mu_{n+\ell} = E\mu_1\mu_{1+\ell} \quad \text{a.s.} \ \square$$

Table 5.1 Kullback Information between exp(L) and exp($\tilde{L}$) under Model A

The upper and lower figures are E(L-$\tilde{L}$) and SD(L-$\tilde{L}$), respectively

σ \ p	.02	.05	.1	.2	.4	.6	.8
.5	.022	.012	.001	.001	.001	.000	.000
	.174	.106	.072	.038	.012	.004	.001
1	.055	.044	.033	.006	.003	.002	.000
	.334	.302	.255	.151	.074	.020	.001
2	.099	.117	.040	.021	.012	.000	.000
	.427	.478	.360	.265	.109	.046	.011
3	.105	.121	.092	.045	.008	.003	.000
	.430	.427	.397	.282	.124	.054	.018
4	.103	.136	.075	.052	.009	.004	.000
	.418	.448	.409	.325	.141	.062	.017
5	.090	.101	.093	.057	.012	.005	.000
	.426	.419	.385	.302	.157	.075	.020
7	.055	.071	.089	.027	.021	.000	.000
	.376	.369	.370	.298	.155	.066	.024
10	.057	.069	.103	.043	.029	.000	.002
	.345	.360	.384	.308	.151	.072	.025
15	.051	.042	.058	.029	.012	.005	.000
	.352	.334	.314	.304	.164	.076	.031

Table 6.1 The AMSE of the Estimators over 100 Samples [a]

Signal	Successive Heights	Points of Change	Est. 1	Est. 2	Est. 3	Est. 4	Given C.P.[+]	$E(\hat{\sigma}_\varepsilon')$ [*]	$SD(\hat{\sigma}_\varepsilon')$
1	0	none	.071(.012)	.059(.008)	.124(.025)	.067(.008)	.05	.943	.166
2	0,1	10	.228(.012)	.219(.011)	.352(.024)	.251(.009)	.1	.949	.196
3	0,3	10	.254(.018)	.241(.017)	.670(.058)	.385(.033)	.1	.930	.180
4	0,5	10	.187(.019)	.185(.019)	.486(.050)	.189(.018)	.1	.970	.165
5	0,1	15	.204(.008)	.197(.008)	.314(.027)	.197(.007)	.1	.961	.162
6	0,3	15	.301(.024)	.272(.019)	.798(.054)	.385(.036)	.1	.944	.204
7	0,2,4	4,10	.370(.017)	.361(.016)	.802(.054)	.610(.040)	.15	.936	.220
8	0,3,0	7,14	.407(.030)	.371(.025)	.947(.071)	.642(.060)	.15	.916	.258
9	0,3,0	5,15	.395(.031)	.380(.027)	.952(.078)	.780(.067)	.15	.913	.222
10	0,4,6	5,15	.356(.022)	.350(.022)	.793(.070)	.440(.031)	.15	.944	.198

[a]: The number in parentheses next to an entry is the estimated standard error for that entry.

[+]: This column is the AMSE of the estimator using the averages of the data points between successive time points of change.

[*]: The estimated standard error of the estimated $E(\hat{\sigma}_\varepsilon')$ is $SD(\hat{\sigma}_\varepsilon')/10$.

Table 6.1 - Continued

11	0,4,6	8,13	.305(.020)	.303(.018)	.600(.043)	.375(.024)	.15	.944	.189
12	0,1,2,3	4,10,16	.348(.023)	.308(.017)	.729(.041)	.493(.032)	.2	.997	.204
13	0,3,6,9	4,10,16	.477(.027)	.466(.025)	.766(.051)	.504(.031)	.2	1.047	.254
14	0,5,10,15	4,10,16	.328(.032)	.314(.031)	.728(.046)	.305(.032)	.2	.908	.185
15	0,1,0,1	4,10,16	.280(.007)	.265(.008)	.365(.026)	.259(.008)	.2	1.014	.180
16	0,3,0,2	4,10,16	.535(.038)	.434(.022)	.904(.052)	.693(.058)	.2	.980	.314
17	0,1,3,4,6	3,7,12,16	.426(.018)	.415(.017)	.907(.081)	.578(.039)	.25	1.014	.228
18	0,3,-3,6,0	3,7,12,16	.378(.022)	.366(.023)	.984(.034)	.364(.022)	.25	.961	.200
19[1]			.509(.021)	.526(.022)	.875(.063)	.729(.042)	1	1.024	.222
20[2]			.935(.024)	.893(.022)	1.311(.065)	2.124(.033)	1	1.339	.268
21[3]			.854(.024)	.836(.023)	1.037(.045)	1.920(.035)	1	1.238	.280

[1]: Signal 19 is the following. $\mu_n = 0.5(n-1)$, $1 \le n \le 20$

[2]: Signal 20 is the following. $\mu_n = n-1$, $1 \le n \le 11$; $\mu_n = 21 - n$, $12 \le n \le 20$

[3]: Signal 21 is the following. $\mu_n = 10-0.1\,(n-11)^2$, $1 \le n \le 20$

END

DATE
FILMED

3-83

DTIC